

# On the Shannon Covers of Certain Irreducible Constrained Systems of Finite Type\*

Akiko Manada and Navin Kashyap

Dept. Mathematics and Statistics

Queen's University

Kingston, ON K7L 3N6, Canada.

Email: {akiko,nkashyap}@mast.queensu.ca

**Abstract**—A construction of Crochemore, Mignosi and Restivo in the automata theory literature gives a presentation of a finite-type constrained system (FTCS) that is deterministic and has a relatively small number of states. This construction is thus a good starting point for determining the minimal deterministic presentation, known as the Shannon cover, of an FTCS. We analyze in detail the Crochemore-Mignosi-Restivo (CMR) construction in the case when the list of forbidden words defining the FTCS is of size at most two. We show that if the FTCS is irreducible, then an irreducible presentation for the system can be easily obtained from the CMR presentation. By studying the follower sets of the states in this irreducible presentation, we are able to explicitly determine the Shannon cover in some cases. In particular, our results show that the CMR construction directly yields the Shannon cover in the case of an irreducible FTCS with exactly one forbidden word, but this is not in general the case for FTCS's with two forbidden words.

## I. INTRODUCTION

In the information theory literature, constrained systems have traditionally arisen in the context of coding for recording systems [4], [5], [6]. These systems, under the tag of regular languages, also form the cornerstone of automata and formal language theory in computer science [3, Chapters 3–4]. More recently, constrained systems have come up naturally in the context of code design for bio-molecular computation (see, for example, the survey paper [1]).

To describe the aim of this paper, we need some basic terminology [5], [6] from the theory of constrained systems. Recall that a *labeled graph*,  $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{L})$ , is a finite directed graph with vertex set  $\mathcal{V}$ , edge set  $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$ , and edge labeling  $\mathcal{L} : \mathcal{E} \rightarrow \Sigma$ , where  $\Sigma$  is a finite alphabet. We will refer to the vertices of  $\mathcal{G}$  as *states*. A *constrained system* (sometimes referred to as a *constraint*),  $\mathcal{S}$  or  $\mathcal{S}(\mathcal{G})$ , is the set of all finite-length sequences (*words*) obtained by reading off the labels along finite paths in a labeled graph  $\mathcal{G}$ . The constrained system  $\mathcal{S}(\mathcal{G})$  is said to be *presented* by  $\mathcal{G}$ ; equivalently,  $\mathcal{G}$  is called a *presentation* of  $\mathcal{S}(\mathcal{G})$ . A presentation,  $\mathcal{G}$ , of a constrained system  $\mathcal{S}$  is said to be *deterministic* if at each state of  $\mathcal{G}$ , the outgoing edges are labeled distinctly. Deterministic presentations of a constrained system  $\mathcal{S}$  are used to derive finite-state encoders for  $\mathcal{S}$  (cf. [6, Chapter 4]).

In general, a given constrained system  $\mathcal{S}$  has many different deterministic presentations. However, in practice it is often desirable to present  $\mathcal{S}$  by a deterministic graph with the smallest possible number of states among all deterministic presentations of the constraint. Such a *minimal* presentation, called the *Shannon cover* of the constraint, can be used to find finite-state encoders with a small number of states which directly translates to low complexity of encoding into the constraint. The goal of this paper is to explicitly determine the Shannon cover of a certain class of constrained systems known as irreducible finite-type constraints.

While even the Shannon cover is not in general unique for an arbitrary constrained system, it does turn out to be unique in the important case of irreducible constrained systems which we now define. A labeled graph  $\mathcal{G}$  with set of states  $\mathcal{V}$  is said to be *irreducible* if for any pair of states  $s, t \in \mathcal{V}$ , there is a directed path in  $\mathcal{G}$  that begins at  $s$  and ends at  $t$ . A constrained system  $\mathcal{S}$  is defined to be *irreducible* if it can be presented by an irreducible graph. Equivalently,  $\mathcal{S}$  is irreducible iff for any pair of words  $\mathbf{u}, \mathbf{w} \in \mathcal{S}$ , there exists  $\mathbf{v} \in \mathcal{S}$  such that the concatenation  $\mathbf{uvw}$  is also in  $\mathcal{S}$ . It is well known [6, p. 57, Theorem 2.12] that the Shannon cover of an irreducible constrained system is unique up to labeled graph isomorphism.

The Shannon cover of an irreducible constrained system  $\mathcal{S}$  can be obtained from an irreducible deterministic presentation,  $\mathcal{G}$ , of  $\mathcal{S}$  by a procedure known as *state merging* [6, Section 2.6]. This procedure is best described in terms of the follower sets of states. The *follower set*,  $F(s)$ , of a state  $s$  in  $\mathcal{G}$  is the set of all finite-length words generated by paths in  $\mathcal{G}$  starting at  $s$ . Two states  $s$  and  $t$  in  $\mathcal{G}$  are said to be *follower-set equivalent* if  $F(s) = F(t)$ . In such a situation, states  $s$  and  $t$  can be *merged* resulting in a new graph  $\mathcal{H}$  obtained by first eliminating all edges emanating from  $t$ , redirecting into  $s$  all remaining edges entering  $t$ , and finally eliminating  $t$ . It is easily verified that the resulting graph  $\mathcal{H}$  is also an irreducible deterministic presentation of  $\mathcal{S}$ . Recursively carrying out the state merging procedure finally results in an irreducible deterministic presentation,  $\mathcal{K}$ , of  $\mathcal{S}$  that is also *follower-separated*, which means that distinct pairs of states in  $\mathcal{K}$  have distinct follower sets. This graph  $\mathcal{K}$  is the Shannon cover of the constraint. In fact, a deterministic presentation of an irreducible constraint is the Shannon cover of the constraint iff it is irreducible and follower-separated.

\*This work was supported in part by a research grant from the Natural Sciences and Engineering Research Council (NSERC) of Canada.

It is clear that the Shannon cover of a constraint  $\mathcal{S}$  is very simply determined if the state merging procedure can be initiated on a presentation of  $\mathcal{S}$  that already has a relatively small number of states. Such a presentation is obtained for a constrained system of finite type (defined below) via a construction of Crochemore, Mignosi and Restivo [2] which has origins in automata theory. The Crochemore-Mignosi-Restivo (CMR) construction is reasonably amenable to analysis, and we use it as the starting point in our search for the Shannon cover of a constrained system of finite type. In fact, Crochemore *et al.* prove in a result [2, Theorem 14] related to the ones in this paper that their construction yields the Shannon cover for a certain type of finite-type constrained system. However, our results do not follow from theirs.

Let  $\Sigma$  be a finite alphabet. We will denote by  $\Sigma^*$  the set of all finite-length sequences (words) over  $\Sigma$ , including the empty word  $\epsilon$ . If  $\mathbf{x} = x_0x_1 \dots x_{\ell-1}$  is a word over  $\Sigma$ , then any of the subsequences  $x_i x_{i+1} \dots x_j$ ,  $0 \leq i \leq j < \ell$ , is called a *subword* of  $\mathbf{x}$ . By convention, the empty word  $\epsilon$  is a subword of any  $\mathbf{x} \in \Sigma^*$ . A *finite-type constrained system (FTCS)* is characterized by a finite set  $\mathcal{F} \subset \Sigma^*$ , and is defined to be the set,  $\mathcal{S}_{\mathcal{F}}$ , of all words  $\mathbf{w} \in \Sigma^*$  such that  $\mathbf{x}$  does not contain as a subword any word in  $\mathcal{F}$ . The finite set  $\mathcal{F}$  is called a *forbidden set*, and its elements are called *forbidden words*. In this paper, we focus mainly on FTCS's with forbidden sets of cardinality at most two. The difficulties involved in extending our analysis further will already be apparent from the cardinality-two case.

The rest of the paper is organized as follows. The CMR construction is described in Section II, and some useful properties of this construction are given in Section III. Sections IV and V study the Shannon covers of FTCS's with one and two forbidden words, respectively. We show there that the CMR construction directly yields the Shannon cover in the case of an irreducible FTCS with exactly one forbidden word, but this is not in general the case for FTCS's with two forbidden words. Most of the results in this paper are stated without proof. Complete proofs of these results will be provided in the full version of the paper.

## II. THE CMR CONSTRUCTION

We fix a finite alphabet  $\Sigma$ , and let  $\mathcal{F} \subset \Sigma^*$  be a non-empty finite set. We assume that  $\mathcal{F}$  is a *non-redundant* collection of words in that no word  $\mathbf{u} \in \mathcal{F}$  is a subword of any  $\mathbf{w} \in \mathcal{F}$ ,  $\mathbf{w} \neq \mathbf{u}$ . Define a labeled graph  $\mathcal{D}_{\mathcal{F}} = (\mathcal{V}, \mathcal{E}, \mathcal{L})$  as follows:

- $\mathcal{V} = \{\mathbf{w} : \mathbf{w} \text{ is a prefix of a word in } \mathcal{F}\}$ . Note that a prefix of a word  $\mathbf{x} = x_0x_1 \dots x_{\ell-1}$  is any of its subwords  $x_0x_1 \dots x_j$  for  $0 \leq j < \ell$ , or the empty word  $\epsilon$ . The states corresponding to  $\mathbf{w} \in \mathcal{F}$  will be called *sink states*, and we will often refer to the state corresponding to the empty word  $\epsilon$  as the *initial state*.
- There are no edges emanating from any sink state  $\mathbf{w} \in \mathcal{F}$ . There are  $|\Sigma|$  edges, all having distinct labels, emanating from each state  $\mathbf{u} \in \mathcal{V} \setminus \mathcal{F}$ . These edges are defined in the following manner: for each  $a \in \Sigma$ ,

if  $\mathbf{ua} \in \mathcal{V}$ , then the edge labeled  $a$  from  $\mathbf{u}$  is a *forward edge* that terminates at the state  $\mathbf{ua}$ ;

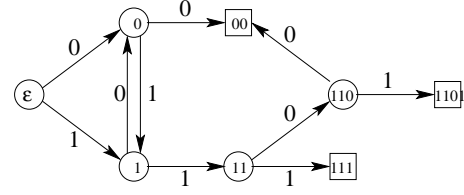


Fig. 1. The CMR automaton  $\mathcal{D}_{\mathcal{F}}$  for  $\mathcal{F} = \{00, 1101, 111\}$  and alphabet  $\Sigma = \{0, 1\}$ . The squares represent the sink states.

if  $\mathbf{ua} \notin \mathcal{V}$ , the edge labeled  $a$  from  $\mathbf{u}$  is a *backward edge* that terminates at the state  $\mathbf{v}$ , where  $\mathbf{v}$  is the longest suffix (incl. the empty word  $\epsilon$ ) of  $\mathbf{ua}$  in  $\mathcal{V}$ .

The graph thus obtained will be referred to as the *CMR automaton* [2]. Figure 1 shows such a graph for  $\mathcal{F} = \{00, 1101, 111\}$  and alphabet  $\Sigma = \{0, 1\}$ .

Let  $\mathcal{G}_{\mathcal{F}}$  be the graph obtained by deleting from  $\mathcal{D}_{\mathcal{F}}$  all sink states and all edges entering sink states. It follows from [2, Theorem 10] that  $\mathcal{G}_{\mathcal{F}}$  is a presentation of the FTCS  $\mathcal{S}_{\mathcal{F}}$  having forbidden set  $\mathcal{F}$ . We will refer to this graph  $\mathcal{G}_{\mathcal{F}}$  as the *CMR presentation* of  $\mathcal{S}_{\mathcal{F}}$ . It is easily seen that both  $\mathcal{D}_{\mathcal{F}}$  and  $\mathcal{G}_{\mathcal{F}}$  are deterministic. The CMR presentation for  $\mathcal{F} = \{00, 1101, 111\}$  is the graph given in Figure 2, without the dotted edges.

By construction, the number of states in  $\mathcal{D}_{\mathcal{F}}$  is at most  $1 + \sum_{\mathbf{w} \in \mathcal{F}} \ell(\mathbf{w})$ , and hence, that in  $\mathcal{G}_{\mathcal{F}}$  is at most  $1 + \sum_{\mathbf{w} \in \mathcal{F}} (\ell(\mathbf{w}) - 1)$ , where  $\ell(\mathbf{w})$  denotes the length of the word  $\mathbf{w}$ . Note that  $1 + \sum_{\mathbf{w} \in \mathcal{F}} (\ell(\mathbf{w}) - 1) \leq |\mathcal{F}| \ell_{\max}$ , where  $\ell_{\max} = \max\{\ell(\mathbf{w}) : \mathbf{w} \in \mathcal{F}\}$ . In comparison, the number of states in the canonical deterministic presentation of  $\mathcal{S}_{\mathcal{F}}$  obtained from the higher edge graph of order  $\ell_{\max}$  of the unconstrained  $\Sigma$ -ary system [5] is  $|\Sigma|^{\ell_{\max}-1}$ , which is typically much larger than  $|\mathcal{F}| \ell_{\max}$ . Thus,  $\mathcal{G}_{\mathcal{F}}$  is in general a better candidate on which to initiate the state merging procedure to construct the Shannon cover than the canonical presentation of  $\mathcal{S}_{\mathcal{F}}$ .

## III. SOME USEFUL PROPERTIES OF $\mathcal{G}_{\mathcal{F}}$

In this section, we give some properties of the CMR presentation  $\mathcal{G}_{\mathcal{F}}$  that will be useful in the subsequent development. We start with the following observation [2, Remark 6(1)], which is an easy consequence of the definitions of  $\mathcal{D}_{\mathcal{F}}$  and  $\mathcal{G}_{\mathcal{F}}$ .

*Lemma 3.1:* For any non-initial state  $\mathbf{u} = u_0u_1 \dots u_{j-1}$ ,  $j > 0$ , in  $\mathcal{D}_{\mathcal{F}}$  or  $\mathcal{G}_{\mathcal{F}}$ , the edges entering  $\mathbf{u}$  all share the same label  $u_{j-1}$ , which is the last symbol of  $\mathbf{u}$ . Hence, a non-initial state in  $\mathcal{D}_{\mathcal{F}}$  or  $\mathcal{G}_{\mathcal{F}}$  has at most one self-loop attached to it.

For the CMR automaton  $\mathcal{D}_{\mathcal{F}}$ , let  $\delta : (\mathcal{V} \setminus \mathcal{F}) \times \Sigma \rightarrow \mathcal{V}$  be the *transition function* defined by setting  $\delta(\mathbf{u}, a)$ ,  $\mathbf{u} \in \mathcal{V}$ ,  $a \in \Sigma$ , to be the state reached by the edge labeled  $a$  emanating from  $\mathbf{u}$ . Note that if the edge labeled  $a$  starting at  $\mathbf{u}$  is a forward edge, then<sup>1</sup>  $\ell(\delta(\mathbf{u}, a)) = \ell(\mathbf{u}) + 1$ , and if it is a backward edge, then  $\ell(\delta(\mathbf{u}, a)) \leq \ell(\mathbf{u})$ .

<sup>1</sup>Since states in  $\mathcal{D}_{\mathcal{F}}$  and  $\mathcal{G}_{\mathcal{F}}$  are identified with words in  $\Sigma^*$ , the notation  $\ell(\mathbf{u})$  for a state  $\mathbf{u}$  simply denotes the length of the word  $\mathbf{u}$ .

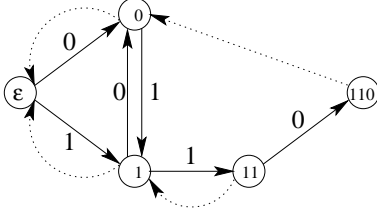


Fig. 2. The CMR presentation  $\mathcal{G}_{\mathcal{F}}$  for  $\mathcal{F} = \{00, 1101, 111\}$  and alphabet  $\Sigma = \{0, 1\}$ . The dotted edges represent the failure function.

Following [2], we will find it convenient to define the notion of a *failure function*,  $f : \mathcal{V} \setminus (\mathcal{F} \cup \{\epsilon\}) \rightarrow \mathcal{V}$ , recursively via

- for each  $a \in \Sigma$ , if  $\delta(\epsilon, a) \in \mathcal{V} \setminus \mathcal{F}$ , then  $f(\delta(\epsilon, a)) = \epsilon$ ;
- for each  $\mathbf{u} \in \mathcal{V} \setminus (\mathcal{F} \cup \{\epsilon\})$  and  $a \in \Sigma$ , if  $\delta(\mathbf{u}, a) \in \mathcal{V} \setminus \mathcal{F}$ , then  $f(\delta(\mathbf{u}, a)) = \delta(f(\mathbf{u}), a)$ .

Note that the failure function is not defined for the initial state and the sink states. The usefulness of the failure function stems from the fact that it helps in efficiently locating the terminal states of the backward edges in  $\mathcal{D}_{\mathcal{F}}$  (and hence in  $\mathcal{G}_{\mathcal{F}}$ ). Indeed, if for some  $\mathbf{u} \in \mathcal{V} \setminus (\mathcal{F} \cup \{\epsilon\})$  and  $a \in \Sigma$ , we have  $\mathbf{u}a \notin \mathcal{V}$ , then  $\delta(\mathbf{u}, a) = \delta(f(\mathbf{u}), a)$ .

The states in the CMR presentation  $\mathcal{G}_{\mathcal{F}}$  simply retain the failure function (as well as the transition function whenever it can be defined) from  $\mathcal{D}_{\mathcal{F}}$ . The dotted edges in Figure 2 represent the failure function for the states in  $\mathcal{G}_{\{00, 1101, 111\}}$ . We will follow this convention of using dotted edges to represent the failure function throughout the paper.

We record in Lemma 3.2 and Proposition 3.3 below some facts about the failure function that we will use in later sections of the paper.

For a state  $\mathbf{u} \neq \epsilon$  in  $\mathcal{G}_{\mathcal{F}}$ , define  $\Delta(\mathbf{u}) = \ell(\mathbf{u}) - \ell(f(\mathbf{u}))$ .

**Lemma 3.2:** Let  $\mathbf{u}, \mathbf{v}$  be non-initial states in  $\mathcal{G}_{\mathcal{F}}$  such that  $\mathbf{u}$  is a prefix of  $\mathbf{v}$ . Then,  $\Delta(\mathbf{u}) \leq \Delta(\mathbf{v})$ .

**Proposition 3.3:** For any state  $\mathbf{u}a$ , with  $\mathbf{u} \in \Sigma^*$  and  $a \in \Sigma$ , in  $\mathcal{G}_{\mathcal{F}}$ , we have  $f(\mathbf{u}a) = \mathbf{u}$  if and only if  $\mathbf{u} = a^t$  for some integer  $t \geq 0$ .

*Remark:* By convention,  $a^0 = \epsilon$ .

Recall from Section I that if  $\mathcal{S}_{\mathcal{F}}$  is an irreducible constrained system, then the Shannon cover of  $\mathcal{S}_{\mathcal{F}}$  is obtained by applying the state merging procedure to an irreducible deterministic presentation of  $\mathcal{S}_{\mathcal{F}}$ . Now,  $\mathcal{G}_{\mathcal{F}}$  is certainly deterministic, but need not always be irreducible. However, it does turn out to be so in most cases, as we shall see in Sections IV and V. The next lemma is a key component in our proofs of irreducibility.

Given a state  $\mathbf{v}$  in  $\mathcal{G}_{\mathcal{F}}$ , let  $N_f(\mathbf{v})$  and  $N_b(\mathbf{v})$  respectively denote the number of forward and backward edges that emanate from  $\mathbf{v}$  in the CMR automaton  $\mathcal{D}_{\mathcal{F}}$ .

**Lemma 3.4:** Let  $l \geq 0$  be an integer such that every state  $\mathbf{u}$  in  $\mathcal{G}_{\mathcal{F}}$  with  $\ell(\mathbf{u}) \leq l$  has a path leading to a distinguished state  $\mathbf{w}$ . If  $\mathbf{v}$  is a state with  $\ell(\mathbf{v}) = l + 1$  such that  $N_b(\mathbf{v}) \geq 1$  and either  $N_f(f(\mathbf{v})) < N_b(\mathbf{v})$  or  $\Delta(\mathbf{v}) \geq 2$  holds, then  $\mathbf{v}$  has a path leading to  $\mathbf{w}$  as well.

The following result is an application of Lemma 3.4.

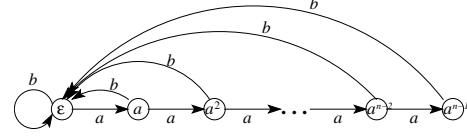


Fig. 3.  $\mathcal{G}_{\mathcal{F}}$  for  $\mathcal{F} = \{a^n\}$

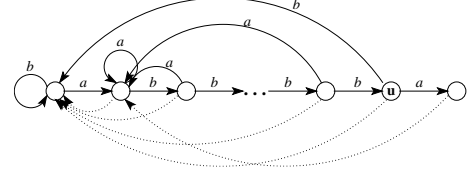


Fig. 4. Initial part of  $\mathcal{G}_{\mathcal{F}}$  for  $\mathcal{F} = \{ab^r ay\}$ ,  $\mathbf{y} \in \Sigma^*$ ,  $1 \leq r \leq n - 2$

**Corollary 3.5:** Let  $|\Sigma| \geq 3$ . If  $N_f(\mathbf{v}) \leq (|\Sigma| - 1)/2$  for all states  $\mathbf{v}$  in  $\mathcal{G}_{\mathcal{F}}$ , then  $\mathcal{G}_{\mathcal{F}}$  is irreducible.

The last lemma in this section gives an important necessary condition for two states in  $\mathcal{G}_{\mathcal{F}}$  to be follower-set equivalent. While it only applies to cases in which all forbidden words have the same length, it is enough for our purposes.

**Lemma 3.6:** Let  $\mathcal{F} \subset \Sigma^*$  be a finite set with the property that all words in  $\mathcal{F}$  have the same length. If  $\mathbf{x}, \mathbf{y}$  are a pair of states in  $\mathcal{G}_{\mathcal{F}}$  that are follower-set equivalent, then  $\ell(\mathbf{x}) = \ell(\mathbf{y})$ .

Note that if  $\mathcal{F}$  consists of exactly one word  $\mathbf{w}$ , then the states of  $\mathcal{G}_{\mathcal{F}}$  are precisely all the distinct proper prefixes of  $\mathbf{w}$ , which are all of different lengths. We thus have

**Corollary 3.7:** If  $|\mathcal{F}| = 1$ , then  $\mathcal{G}_{\mathcal{F}}$  is follower-separated.

We investigate the case of forbidden sets of cardinality one in more detail in the next section.

#### IV. THE CASE OF ONE FORBIDDEN WORD

When the forbidden set consists of exactly one forbidden word, we have a complete and concise result.

**Theorem 4.1:** Let  $\mathcal{F} = \{\mathbf{w}\}$  for some  $\mathbf{w} \in \Sigma^n$ ,  $n \geq 1$ . If  $\mathcal{S}_{\mathcal{F}}$  is irreducible, then  $\mathcal{G}_{\mathcal{F}}$  is the Shannon cover of  $\mathcal{S}_{\mathcal{F}}$ .

*Proof:* We have to show that  $\mathcal{G}_{\mathcal{F}}$  is irreducible and follower-separated. By dint of Corollary 3.7, it is enough to show that  $\mathcal{G}_{\mathcal{F}}$  is irreducible whenever  $\mathcal{S}_{\mathcal{F}}$  is. In fact, it is enough to consider the case of a binary alphabet  $\Sigma$ , since Corollary 3.5 disposes of the  $|\Sigma| \geq 3$  case.

So, let  $\Sigma = \{a, b\}$ . Without loss of generality (WLOG), we may assume that the forbidden word  $\mathbf{w}$  begins with the symbol  $a$ . Note that if  $\mathbf{w} = ab^{n-1}$ , then  $\mathcal{S}_{\mathcal{F}}$  is not irreducible, since  $a, b^{n-1} \in \mathcal{S}_{\mathcal{F}}$ , but there is no word  $\mathbf{x} \in \{a, b\}^*$  such that  $axb^{n-1} \in \mathcal{S}_{\mathcal{F}}$ . Similarly,  $\mathcal{S}_{\mathcal{F}}$  is not irreducible when  $\mathbf{w} = a^{n-1}b$ . For all other words  $\mathbf{w}$ , as we shall see,  $\mathcal{G}_{\mathcal{F}}$  (and hence  $\mathcal{S}_{\mathcal{F}}$ ) is irreducible.

We start with  $\mathbf{w} = a^n$ . It is easily seen that in this case,  $\mathcal{G}_{\mathcal{F}}$  is as in Figure 3, which is seen to be irreducible by inspection.

Next, let  $\mathbf{w} = ab^r ay$ , with  $\mathbf{y} \in \Sigma^*$  and  $1 \leq r \leq n - 2$ . We will show that all states in  $\mathcal{G}_{\mathcal{F}}$  have a path going to the initial state  $\epsilon$ . Figure 4 shows the subgraph of  $\mathcal{G}_{\mathcal{F}}$  containing the states from  $\epsilon$  up to  $\mathbf{u} = ab^r$ . Since there is an edge from

$\mathbf{u}$  to  $\epsilon$ , we see that there is a path starting from any state between  $\epsilon$  and  $\mathbf{u}$  that goes back to the initial state  $\epsilon$ . To see that this is also the case for states beyond  $\mathbf{u}$ , we use Lemma 3.4. From Figure 4, we note that  $f(\mathbf{ua}) = a$ . Thus,  $\Delta(\mathbf{ua}) = \ell(\mathbf{ua}) - \ell(f(\mathbf{ua})) = \ell(\mathbf{u}) \geq 2$ . Therefore, by Lemma 3.4, there is a path from  $\mathbf{ua}$  to the initial state  $\epsilon$ . For states  $\mathbf{v}$  with  $\ell(\mathbf{v}) > \ell(\mathbf{ua})$ , by Lemma 3.2, we have  $\Delta(\mathbf{v}) \geq 2$  as well. So, repeated application of Lemma 3.4 shows that any such  $\mathbf{v}$  also has a path going to the initial state  $\epsilon$ . Thus,  $\mathcal{G}_{\mathcal{F}}$  is irreducible.

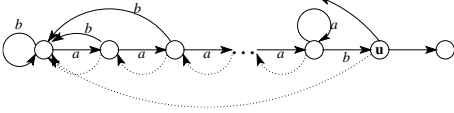


Fig. 5. Initial part of  $\mathcal{G}_{\mathcal{F}}$  for  $\mathcal{F} = \{a^r by\}$ ,  $y \in \Sigma^*$ ,  $2 \leq r \leq n-2$

Finally, let  $\mathbf{w} = a^r by$ , with  $y \in \Sigma^*$  and  $2 \leq r \leq n-2$ . Figure 5 shows the subgraph of  $\mathcal{G}_{\mathcal{F}}$  containing the states from  $\epsilon$  up to  $\mathbf{u} = a^r b$ . We shall show first that there is a path to the initial state from the state  $\mathbf{u}$ . Since  $f(\mathbf{u}) = \epsilon$ , the backward edge from  $\mathbf{u}$ , if labeled  $a$ , goes to the state  $a$ , and if labeled  $b$ , goes to the initial state. But since there is an edge from the state  $a$  to the initial state, there is always a path from  $\mathbf{u}$  to  $\epsilon$ . In addition since  $\Delta(\mathbf{u}) = \ell(\mathbf{u}) \geq 2$ , we also have  $\Delta(\mathbf{v}) \geq 2$  for states  $\mathbf{v}$  with  $\ell(\mathbf{v}) > \ell(\mathbf{u})$ , by Lemma 3.2. Thus, as before, repeated application of Lemma 3.4 shows that any such  $\mathbf{v}$  also has a path going to the initial state  $\epsilon$ , proving that  $\mathcal{G}_{\mathcal{F}}$  is irreducible. This completes the proof of the theorem. ■

## V. THE CASE OF TWO FORBIDDEN WORDS

When the forbidden set consists of more than one word, the analysis gets a lot more complicated. The intricacies of the analysis become evident even in the case of forbidden sets of size two. In this section, we consider forbidden sets  $\mathcal{F} = \{\mathbf{w}_1, \mathbf{w}_2\} \subset \Sigma^*$ ,  $\mathbf{w}_1 \neq \mathbf{w}_2$ , with  $\ell(\mathbf{w}_1) = \ell(\mathbf{w}_2)$ . Furthermore, we will only present results for the case when  $|\Sigma| \geq 3$ , as the results for the binary alphabet do not have simple statements in many cases. For example, when  $|\Sigma| \geq 3$ ,  $\mathcal{G}_{\mathcal{F}}$  is itself irreducible (Theorem 5.1), while in the binary case, we sometimes have to pass to a (proper) subgraph of  $\mathcal{G}_{\mathcal{F}}$  to obtain an irreducible presentation of  $\mathcal{S}_{\mathcal{F}}$ .

So, for the rest of this section, we will assume a finite alphabet  $\Sigma$  with  $|\Sigma| \geq 3$ , and a subset  $\mathcal{F} \subset \Sigma^*$  consisting of two distinct equal-length words,  $\mathbf{w}_1$  and  $\mathbf{w}_2$ . We set  $\mathbf{p}$  to be the longest common prefix (including the empty word  $\epsilon$ ) of  $\mathbf{w}_1$  and  $\mathbf{w}_2$ . Note that by construction,  $\mathbf{p}$  is the only state in  $\mathcal{G}_{\mathcal{F}}$  with two forward edges; all other states have at most one forward edge. We say that  $\mathcal{G}_{\mathcal{F}}$  forks at  $\mathbf{p}$ , as  $\mathcal{G}_{\mathcal{F}}$  forks into two branches “downstream” from  $\mathbf{p}$ , as depicted in Figure 6.

**Theorem 5.1:** If  $\mathcal{S}_{\mathcal{F}}$  is an irreducible FTCS, then  $\mathcal{G}_{\mathcal{F}}$  is irreducible as a directed graph.

*Sketch of proof:* If  $|\Sigma| \geq 5$ , then Corollary 3.5 gives us the irreducibility of  $\mathcal{G}_{\mathcal{F}}$ . If  $|\Sigma| = 4$ , then proving the irreducibility of  $\mathcal{G}_{\mathcal{F}}$  is still a relatively easy application of Lemma 3.4. We skip the details.

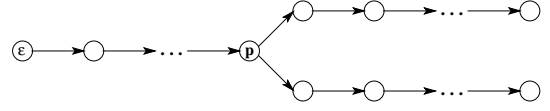


Fig. 6. A typical  $\mathcal{G}_{\mathcal{F}}$  for  $|\mathcal{F}| = 2$ . Only forward edges are shown.

So, suppose that  $\Sigma = \{a, b, c\}$ , and let  $\ell(\mathbf{w}_1) = \ell(\mathbf{w}_2) = n$ , and  $\ell(\mathbf{p}) = \rho$ . We will show that each state in  $\mathcal{G}_{\mathcal{F}}$  has a path leading to the initial state  $\epsilon$ . We divide the proof into three cases: (a)  $\rho = 0$ ; (b)  $1 \leq \rho \leq n-2$ ; and (c)  $\rho = n-1$ . We give complete proofs for the first and last cases as illustrations, but skip the proof for Case (b).

**Case (a):  $\rho = 0$ .** Here, the graph  $\mathcal{G}_{\mathcal{F}}$  forks at the initial state itself. WLOG, the two forward edges from  $\epsilon$  are labeled  $a$  and  $b$ , respectively. It is enough to show that the states  $a$  and  $b$  each have a path going to the initial state. Indeed, if this can be shown, then it follows from Lemma 3.4 that the states  $\mathbf{v}$  with  $\ell(\mathbf{v}) \geq 2$  also have paths going to the initial state, as these states satisfy the conditions of that lemma.

Suppose that one of the states  $a$  and  $b$  has an edge going to the initial state  $\epsilon$ . WLOG, let this state be  $a$ . The state  $b$  has two backward edges, of which at most one can be a self-loop by Lemma 3.1. Thus, the other edge goes either to  $\epsilon$  or to  $a$ . In any case,  $b$  also has a path going to the initial state.

We are left to deal with the situation when neither  $a$  nor  $b$  has an edge going to the initial state. In this situation, both  $a$  and  $b$  have self-loops,  $a$  has an edge going to  $b$ , and  $b$  has an edge terminating at  $a$ . It is straightforward to see that this can happen only if  $\mathcal{F} = \{acx, bcy\}$  for some  $x, y \in \Sigma^*$ , in which case the initial part of the graph  $\mathcal{G}_{\mathcal{F}}$  is as in Figure 7.

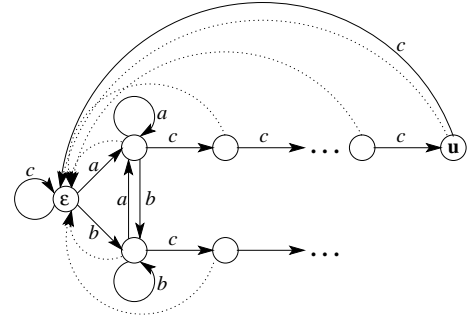


Fig. 7. Initial part of  $\mathcal{G}_{\mathcal{F}}$  for the case  $\mathcal{F} = \{acx, bcy\}$  for some  $x, y \in \Sigma^*$ .

Now, if  $x = y = c^{n-2}$ , then we have  $\mathcal{F} = \{ac^{n-1}, bc^{n-1}\}$ , in which case  $\mathcal{S}_{\mathcal{F}}$  is not irreducible, since  $a, c^{n-1} \in \mathcal{S}_{\mathcal{F}}$ , but there can be no  $z \in \{a, b, c\}^*$  such that  $azc^{n-1} \in \mathcal{S}_{\mathcal{F}}$ . So, assuming WLOG that  $x \neq c^{n-2}$ , there is a largest  $r < n-1$  for which  $ac^r$  is a state; let  $\mathbf{u}$  denote the state  $ac^r$  corresponding to this largest  $r$ . As shown in Figure 7,  $\delta(\mathbf{u}, c) = \epsilon$ . Thus, the states  $a$  and  $b$  both have paths leading to the initial state, as can be verified from the figure.

We have thus proved that  $\mathcal{G}_{\mathcal{F}}$  is irreducible whenever  $\rho = 0$ .

**Case (c):  $\rho = n-1$ .** If  $\mathbf{p} \notin \{a^{n-1}, b^{n-1}, c^{n-1}\}$ , then  $\mathcal{G}_{\mathcal{F}}$  can easily be shown to be irreducible using Lemma 3.4. So, assume WLOG that  $\mathcal{F}$  is either  $\{a^{n-1}b, a^{n-1}c\}$  or  $\{a^n, a^{n-1}b\}$ .

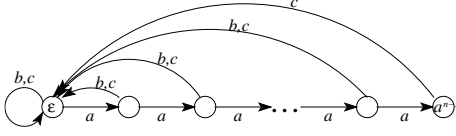


Fig. 8.  $\mathcal{G}_F$  for  $\mathcal{F} = \{a^n, a^{n-1}b\}$  and  $\Sigma = \{a, b, c\}$ .

Note, however, that when  $\mathcal{F} = \{a^{n-1}b, a^{n-1}c\}$ ,  $\mathcal{S}_F$  is not irreducible. When  $\mathcal{F} = \{a^n, a^{n-1}b\}$ ,  $\mathcal{G}_F$  is as shown in Figure 8, and is clearly irreducible. ■

Thus, by applying the state merging procedure to  $\mathcal{G}_F$ , we can obtain the Shannon cover of  $\mathcal{S}_F$ . To do this, we must of course identify the states in  $\mathcal{G}_F$  that are follower-set equivalent. This also turns out to be a non-trivial task, and we are at present able to give a complete solution only in the special case when  $\mathcal{F} = \{a^n, ax\}$  for some  $x \in \Sigma^{n-1}$ ,  $x \neq a^{n-1}$ .

In the following exposition, we set  $\mathbf{z}_1 = a^n$ , and  $\mathbf{z}_2 = ax$  for some  $x \in \Sigma^{n-1}$ ,  $x \neq a^{n-1}$ . Note that we can parse  $\mathbf{z}_2$  uniquely as

$$\mathbf{z}_2 = a^{x_1} \beta^{(1)} a^{x_2} \beta^{(2)} \dots a^{x_{q-1}} \beta^{(q-1)} a^{x_q}$$

for some integer  $q \geq 2$ , where  $x_1, x_2, \dots, x_{q-1}$  are positive integers,  $x_q$  is a non-negative integer, and  $\beta^{(j)} \in (\Sigma \setminus \{a\})^*$  for  $j = 1, 2, \dots, q-1$ . WLOG, we assume that  $\beta^{(1)}$  begins with the symbol  $b \neq a$ .

Figure 9 shows the generic structure of  $\mathcal{G}_F$  for  $\mathcal{F} = \{\mathbf{z}_1, \mathbf{z}_2\}$ . From Theorem 5.1, we know that  $\mathcal{G}_F$  is irreducible. And as stated in the next result, this presentation is also follower-separated when  $x_1 \geq x_q$ .

**Theorem 5.2:** Let  $\mathcal{F} = \{\mathbf{z}_1, \mathbf{z}_2\}$ . If  $x_1 \geq x_q$ ,  $\mathcal{G}_F$  is follower-separated, and hence is the Shannon cover of  $\mathcal{S}_F$ .

To state the corresponding result for the case when  $x_1 < x_q$ , we need additional notation and terminology. For  $j = 1, 2, \dots, q-1$ , we define certain distinguished prefixes of  $\mathbf{z}_2$ ,

$$\mathbf{p}_j = a^{x_1} \beta^{(1)} a^{x_2} \beta^{(2)} \dots a^{x_j} \beta^{(j)} a,$$

and set  $\mathbf{p}_0 = a$ . The states  $\mathbf{p}_j$  in  $\mathcal{G}_F$  satisfy the following property.

**Lemma 5.3:** For  $j > 0$ ,  $f(\mathbf{p}_j) = \mathbf{p}_k$  for some  $k < j$ .

Thus, we can define the set of indices

$$\text{Ind}_f = \{k : f^r(\mathbf{p}_{q-1}) = \mathbf{p}_k \text{ for some } r \geq 1\},$$

where  $f^r(\cdot)$  denotes the  $r$ th iterate of the failure function  $f$ . Note that  $0 \in \text{Ind}_f$ , since some iterate of the failure function will eventually take  $\mathbf{p}_{q-1}$  to  $\mathbf{p}_0 = a$ .

For  $l = 0, 1, \dots, n-1$ , let us define  $\Lambda_l$  to be the set of all states  $\mathbf{u}$  in  $\mathcal{G}_F$  such that  $\ell(\mathbf{u}) = l$ . Thus,  $|\Lambda_l| = 1$  if  $l \leq x_1$ , and  $|\Lambda_l| = 2$  if  $l > x_1$ . We will often say that the states in  $\Lambda_l$  are at level  $l$  in  $\mathcal{G}_F$ . Recall from Lemma 3.6 that two states in  $\mathcal{G}_F$  are follower-set equivalent only if they are at the same level. We can now state the result for  $x_1 < x_q$ .

**Theorem 5.4:** Let  $\mathcal{F} = \{\mathbf{z}_1, \mathbf{z}_2\}$ . If  $x_1 < x_q$ , define

$$\mathcal{X} = \{x_{k+1} : k \in \text{Ind}_f, k > 0, \text{ and } x_1 \leq x_{k+1} < x_q\}.$$

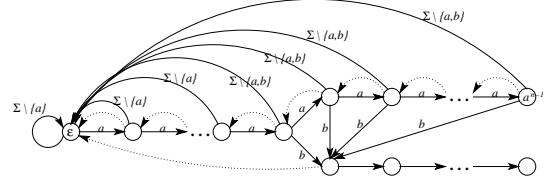


Fig. 9. Typical  $\mathcal{G}_F$  for the case  $\mathcal{F} = \{a^n, ax\}$  for some  $x \in \Sigma^{n-1}$ .

If  $\mathcal{X} \neq \emptyset$ , set  $x^* = \max \mathcal{X}$ ; else, set  $x^* = x_1 - 1$ . Then, the states at level  $l \geq \ell(\mathbf{p}_1)$  are follower-set equivalent iff  $l \geq \ell(\mathbf{p}_{q-1}) + x^*$ . Consequently, the Shannon cover of  $\mathcal{S}_F$  is obtained from  $\mathcal{G}_F$  by merging the pair of states at each level  $l \geq \ell(\mathbf{p}_{q-1}) + x^*$ .

Theorems 5.2 and 5.4 completely specify the Shannon cover in the case of  $\mathcal{F} = \{\mathbf{z}_1, \mathbf{z}_2\}$ . As a direct corollary of these theorems, we have the following result.

**Corollary 5.5:** For  $\mathcal{F} = \{\mathbf{z}_1, \mathbf{z}_2\}$ , the number of states,  $\nu_F$ , in the Shannon cover of  $\mathcal{S}_F$  is given by

$$\nu_F = \begin{cases} 2n - x_1 - 1 & \text{if } x_1 \geq x_q \\ 2n - x_1 - (x_q - x^*) & \text{if } x_1 < x_q \end{cases}$$

Generalizing the above result to arbitrary  $\mathcal{F}$ 's of size two is by no means easy. We do have the following simple bound in the case of alphabets of size at least three, but finding tighter bounds or exact results remains an open problem.

**Theorem 5.6:** Let  $\mathcal{F} = \{\mathbf{w}_1, \mathbf{w}_2\}$ , for some  $\mathbf{w}_1, \mathbf{w}_2 \in \Sigma^n$ ,  $\mathbf{w}_1 \neq \mathbf{w}_2$ . Define  $\rho$  and  $\sigma$  to be the lengths of the longest common prefix and the longest common suffix, respectively, of  $\mathbf{w}_1$  and  $\mathbf{w}_2$ . Then, the number of states,  $\nu_F$ , in the Shannon cover of  $\mathcal{S}_F$  can be bounded as

$$2n - \rho - \sigma - 1 \leq \nu_F \leq 2n - \rho - 1$$

The results of Theorems 5.1, 5.2, 5.4 and 5.6 can be extended, upon appropriate modification, to binary alphabets as well. But as the statements in the binary case are a lot more dense, we do not present them in this paper. The results for the binary alphabet, as well as complete proofs of the results given here, will be published in the full version of this paper.

## REFERENCES

- [1] A. Brennenman and A.E. Condon, "Strand design for bio-molecular computation," *Theoretical Computer Science*, vol. 287:1, pp. 39–58, 2002.
- [2] M. Crochemore, F. Mignosi and A. Restivo, "Automata and forbidden words," *Inform. Proc. Lett.*, vol. 67, pp. 111–117, 1998.
- [3] J.E. Hopcroft, R. Motwani and J.D. Ullman, *Introduction to Automata Theory, Languages, and Computation*, 2nd ed., Addison-Wesley, 2001.
- [4] K.A.S. Immink, *Codes for Mass Data Storage Systems*, 2nd ed., Rotterdam, The Netherlands: Shannon Foundation Publishers, 2004.
- [5] B.H. Marcus, R.M. Roth and P.H. Siegel, "Constrained Systems and Coding for Recording Channels," in *Handbook of Coding Theory*, R. Brualdi, C. Huffman and V. Pless, Eds., Amsterdam, The Netherlands: Elsevier, 1998.
- [6] B.H. Marcus, R.M. Roth and P.H. Siegel, *An Introduction to Coding for Constrained Systems*, unpublished lecture notes. Available at <http://www.cs.technion.ac.il/~ronny/constrained.html>.